主観視点画像を用いた探索問題における変換画像の同時学習によるICMの改良

主観視点画像を用いた探索問題における変換画像の同時学習による ICM の改良
Improvement of ICM by Simultaneous Learning of Transformed Images in Exploration Problem Using First Person View Images

飯岡 徹人 *
Tetsuto Iioka
法政大学 情報科学研究科
Email: tetsuto.iioka.7v@stu.hosei.ac.jp

***Abstract*—Reinforcement learning is a powerful method for learning policy, but the policy cannot be learned well in environments where external rewards are sparse. To solve this problem, Intrinsic Curiosity Module (ICM) was proposed. ICM is a method of generating an internal reward that encourages searching for a state that is unlikely to predict the next state; for this purpose, concurrently with the training of the policy, ICM trains a network that predicts the next state by receiving the current state and action. A learning method using ICM is capable of learning policy by using screen images as input states in a sparse reward environment of video games such as Doom. This paper proposes two methods of improving the learning speed by using images that are transformed at the same time when training a network that generates internal reward using images. This paper experimentally compares the learning performance of the proposed methods with that of ICM in the Doom target search environment. The result of the first experiment showed that the method using images whose colors were converted in a negative/positive way was faster in learning policy than the conventional ICM, but the method using images that were mirror-reversed was slower in a map with a specific texture given for the walls of each room. The result of the second experiment showed that the conventional ICM was superior in the generalization ability to the two proposed methods. The result of the third experiment showed that the two proposed methods were faster in learning policy than the conventional ICM in a map with random textures for the walls of the rooms.**

1. はじめに

強化学習 [1] とは，機械学習の一種であり，定められた目標を達成するのに最適な行動の方策を学習するための手法である．近年では，強化学習にニューラルネットワークを組み合わせた深層強化学習の手法が次々と提案されている．DQN[2] や A3C[3] を始めとするこれらの手法は，従来の強化学習の手法では学習できない複雑な環境でも学習を行うことができる．しかし，強化学習のアルゴリズムには本質的な弱点があり，方策を獲得するための訓練を行うエージェントに対して，環境からの報酬が与えられる機会がほとんどない問題，例えば目標となる状態に到達した場合のみ報酬が与えられるような問題では，エージェントの行動が目標を達成するために適している行動かどうかの評価を行うことができず，方策を上手く学習することができない．

このような環境でも方策を学習できる手法として，Pathak らによって Intrinsic Curiosity Module (ICM)[4] を用いた強化学習が提案されている．ICM は強化学習によって方策を学習する際，環境から与えられる報酬とは別に，探索の進んでいない状態へと向かう行動をエージェントに促すことができる内部報酬を生成する．

本研究では，ICM のネットワークを訓練する際に，環境から与えられた状態とエージェントが選択した行動だけでなく，それらに特定の変換処理を加えたものを用いることで強化学習の学習速度を向上させる手法を提案する．

2. 関連研究

代表的な強化学習の手法として Q 学習がある．Q 学習で訓練を行うエージェントは，現在の状態 s において選択可能な行動 a の価値 $Q(s,a)$ に基づいて次の行動を決定する．この価値関数を Q 値と呼ぶ．

Q 学習における Q 値のようなテーブル関数では，入力される状態の次元数が高くなるとコンピュータのメモリ容量が対応できなくなるので，ニューラルネットワークで価値関数を近似する手法が提案されている．強化学習の手法にニューラルネットワークを組み合わせる手法の代表的なものが，Deep Q-Network (DQN) である．DQN では，Q 学習における Q 値にあたる最適行動価値関数を畳み込みニューラルネットワークで近似している．

Asynchronous Advantage Actor-Critic (A3C) は，DQN に並列計算の考えを加えて更に発展させた強化学習の手法である．Advantage とは，ネットワークのパラメータを更新する際に，現在の状態に対する報酬だけでなく，数ステップ行動するまでにエージェントが受け取る報酬も用いるという考え方である．Actor-Critic という学習手法では，DQN のように価値関数を使用せず，ネットワークが行動を出力する Actor 部分と，報酬を出力する Critic 部分に分かれている．A3C は探索を行う際に，マルチスレッド環境において，それぞれのスレッドの環境の中で独自にエージェントを訓練して経験を集める．各スレッドの環境はそれぞれ独立した Advantage-Actor-Critic のネットワークを持っている．更に，すべての環境で共有しているネットワークであるパラメータサーバが存在する．各 Advantage-Actor-Critic のネットワークは一定ステップごとか，探索が終端に達したとき，集めた経験から方策を最適化するためにネットワークのパラメータを更新する勾配を求める．求められた勾配はパラメータサーバに送られ，パラメータサーバのネットワークの重みが更新される．勾配を渡したスレッドの環境は，パラメータサーバの重みを自身のネットワークにコピーして，探索を継続する．これを各スレッドの環

* Supervisor: Prof. Hiroshi Hosobe

境が非同期的に実施することで，効率的に学習を行うことができる．

機械学習の訓練データに対して加工を行うことによってデータ数を増やすことをデータ拡張という．浅川の論文 [5] で解説されているように，画像データの認識や分類を行うためのモデルの訓練において，学習精度を向上させるなどの目的で，反転，移動，拡大縮小したデータを用いることがある．本研究で行う，環境から受け取った状態に対する変換処理も，ICM の訓練のためのデータ拡張である．

3. 準備

3.1. 強化学習

強化学習エージェントは環境から現在の状態 s_t を受け取り，現在の方策 $\pi(s_t;\theta_P)$ に基づいて行動 a_t を選択し，環境へ返す．θ_P は方策を決定付けるパラメータであり，強化学習の最適行動価値関数にニューラルネットワークを用いている場合は，ネットワークの重みが該当する．環境はエージェントから行動 a_t を受け取り，状態 s_t において行動 a_t を取ることで遷移する状態 s_{t+1} と，s_{t+1} における報酬 r_t をエージェントへ返す．

エージェントは探索が終端に達するまでの一連の行動によって受け取ることができる報酬の総和の期待値を最大化するように，方策 $\pi(s_t;\theta_P)$ のパラメータ θ_P を更新する．これは，以下のように表される．

$$\max_{\theta_P} E_{\pi(s_t;\theta_P)}\left[\sum_t r_t\right]$$

パラメータの更新は，ニューラルネットワークを用いる深層強化学習の場合は誤差逆伝搬法によって行われる．

以上の手順を 1 ステップとし，これを繰り返すことでエージェントは最適な方策を学習する．

強化学習をするエージェントは環境から報酬を受け取らない限り，方策を更新することができない．そのため，目標となる状態へエージェントが到達するまでに多くの行動を必要とするような環境では，強化学習エージェントは上手く方策を学習することができない．

3.2. Intrinsic Curiosity Module (ICM)

ICM は，逆モデルと順モデルの 2 つのニューラルネットワークのモデルからなる．逆モデルは，2 つのニューラルネットワークのサブモジュールで構成されている．1 つ目は，現在の状態 s_t と次の状態 s_{t+1} を入力として受け取り，特徴量 $\phi(s_t)$ と $\phi(s_{t+1})$ を出力するサブモジュールである．2 つ目は，上述のサブモジュールで出力された特徴量 $\phi(s_t)$ と $\phi(s_{t+1})$ を入力として受け取り，状態 s_t から次の状態 s_{t+1} へ遷移する際にエージェントが取った行動 a_t の予測 $\hat{a}_t$ を出力するサブモジュールである．このモデルを

$$\hat{a}_t = g(s_t, s_{t+1};\theta_I)$$

と定義すると，そのパラメータ θ_I はエージェントの取った行動の予測 $\hat{a}_t$ と実際に取った行動 a_t の誤差を表す損失関数 $L_I(\hat{a}_t, a_t)$ を最小化するように学習される．

$$\min_{\theta_I} L_I(\hat{a}_t, a_t)$$

このモデルを訓練することによって，特徴量 $\phi(s_t)$，$\phi(s_{t+1})$ は状態 s_t，s_{t+1} のうち，エージェントの行動に影響を与える特徴，またはエージェントの行動によって影響を及ぼされる特徴，すなわちエージェントの学習に関係のある特徴のみを抽出したものになっていく．

順モデルは，現在エージェントが置かれている状態 s_t の特徴量 $\phi(s_t)$ とその状態でエージェントが選択した行動 a_t を入力として受け取り，状態 s_t で行動 a_t を取ることで遷移した状態 s_{t+1} の特徴量 $\phi(s_{t+1})$ の予測 $\hat{\phi}(s_{t+1})$ を出力する．このモデルを

$$\hat{\phi}(s_{t+1}) = f(\phi(s_t), a_t;\theta_F)$$

と定義すると，そのパラメータ θ_F は行動後に遷移した状態 s_{t+1} の特徴量の予測 $\hat{\phi}(s_{t+1})$ と，実際の状態 s_{t+1} の特徴量 $\phi(s_{t+1})$ との誤差を表す損失関数

$$L_F(\phi(s_{t+1}), \hat{\phi}(s_{t+1})) = \frac{1}{2}||\hat{\phi}(s_{t+1}) - \phi(s_{t+1})||_2^2$$

を最小化するように学習される．

ここで，好奇心に基づいた内部報酬 r_t^i は以下のように定義される．

$$r_t^i = \frac{\eta}{2}||\hat{\phi}(s_{t+1}) - \phi(s_{t+1})||_2^2$$

$\eta > 0$ は倍率定数である．この内部報酬は，順モデルの出力する状態の特徴量の予測 $\hat{\phi}(s_{t+1})$ が実際の特徴量 $\phi(s_{t+1})$ から外れるほど，つまり状態 s_{t+1} に関するエージェントの学習が進んでいないほど大きくなるようになっている．

ICM を組み合わせた強化学習では，順モデルによって生成される内部報酬 r_t^i と，環境から与えられる報酬 r_t^e の和

$$r_t = r_t^i + r_t^e$$

を方策 $\pi(s_t;\theta_P)$ のパラメータの更新の際に使用する．順モデルにとって予測することが難しい状態である程，生成される内部報酬は大きくなるため，ICM を組み合わせた強化学習のエージェントは探索が進んでいない状態へと遷移する行動を積極的に選択するようになる．

ICM を組み合わせた強化学習のエージェントが方策を最適化するということは，以下の関数を最小化することである．

$$\min_{\theta_P,\theta_I,\theta_F}\left[-\lambda E_{\pi(s_t;\theta_P)}\left[\sum_t r_t\right] + (1-\beta)L_I + \beta L_F\right]$$

$0 \leq \beta \leq 1$ は順モデルに対する逆モデルの重み付けをする定数である．$\lambda > 0$ は内部報酬の学習に対する方策の重要性を重み付けする定数である．

4. 提案手法

本研究では，Doom のようなゲーム画面を状態としてエージェントに返す環境において，ICM のニューラルネットワークを訓練する際に，エージェントが状態として受け取ったピクセル画像と，その状態でエージェントが取った行動だけでなく，その状態と行動にそれぞれ特定の変換処理を加えたものも同時に訓練に用いる手法を提案する．変換処理は，変換後の状態が環境中の別の状態に近似できうるものを使用する．提案手法として，状態である画像と行動を鏡写しのように左右反転させる変換と，画像のピクセルの色のネガティブ，ポジティブを反転させる変換の 2 つの手法を評価する．

本研究の提案手法では，まず従来の ICM と同様に，逆モデルに状態 s_t と s_{t+1} を入力して，行動の予測 $\hat{a}_t$ を出力し，実際にエージェントが取った行動 a_t との誤差

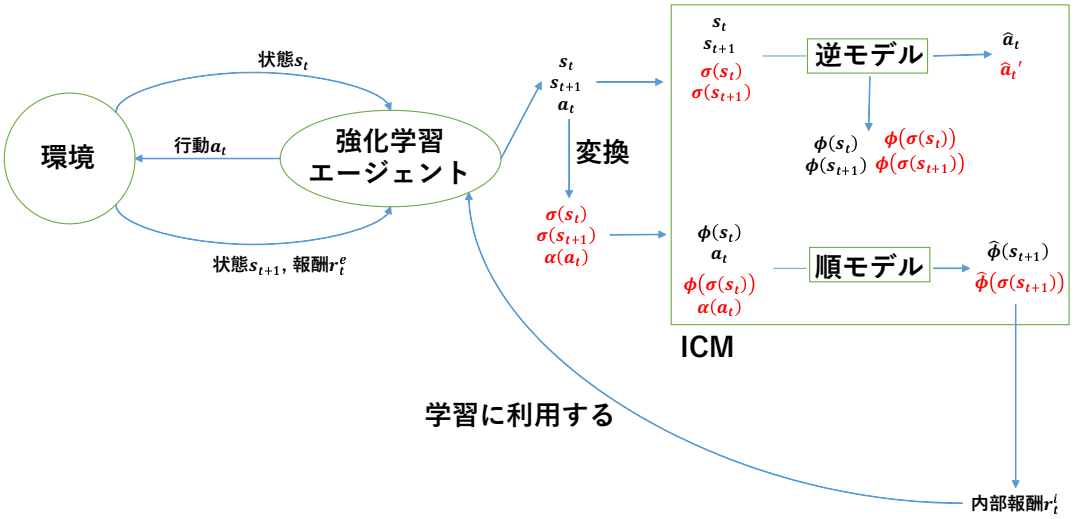

図 1. 変換画像を同時学習する ICM を用いた強化学習

からパラメータ θ_I を更新する．その後，s_t，a_t，s_{t+1} に変換処理を加えた $\sigma(s_t)$，$\alpha(a_t)$，$\sigma(s_{t+1})$ を用いてもう一度逆モデルを訓練する．すなわち，逆モデルに $\sigma(s_t)$ と $\sigma(s_{t+1})$ を入力して，エージェントの取った行動の予測である $\hat{a}_t'$ を出力し，$\alpha(a_t)$ との誤差を最小化するようにパラメータ θ_I を更新する．

順モデルも同様に，まず従来の ICM と同じく逆モデルによって抽出された状態の特徴量 $\phi(s_t)$ とエージェントの取った行動 a_t を入力として受け取り，状態の特徴量の予測 $\hat{\phi}(s_{t+1})$ を出力して，実際の状態の特徴量 $\phi(s_{t+1})$ との誤差からパラメータ θ_F を更新する．次に，逆モデルに $\sigma(s_t)$ を入力することで抽出される，変換された状態の特徴量 $\phi(\sigma(s_t))$ と変換された行動 $\alpha(a_t)$ を入力として受け取り，出力した特徴量の予測 $\hat{\phi}(\sigma(s_{t+1}))$ と実際の特徴量 $\phi(\sigma(s_{t+1}))$ との誤差からパラメータ θ_F を更新する．

ICM によって生成されエージェントへと渡される内部報酬は，変換処理を施していない状態と行動 (s_t, a_t, s_{t+1}) を ICM のネットワークに訓練させて生成された報酬と，変換処理を施した状態と行動 $(\sigma(s_t), \alpha(a_t), \sigma(s_{t+1}))$ を ICM に訓練させて生成された報酬の平均とする．変換画像を同時学習する ICM を用いた強化学習システムの全体図は図 1 のようになる．

提案手法を用いることによって，逆モデルはエージェントが 1 ステップ訓練するごとに 2 ステップ分の訓練が可能になり，より早い段階でエージェントの学習に関係のある状態の特徴量を抽出することができるようになることが期待される．順モデルではエージェントが探索して ICM の訓練を行った状態と行動 (s_t, a_t) に対する内部報酬が以降の訓練で減少することになるが，それと同時に変換処理によって近似された別の状態と行動 $(\sigma(s_t), \alpha(a_t))$ に対する内部報酬も減少する．このため，エージェントが行動を選択する際に従来の ICM と異なる優先基準で行動を選択するので，探索のパターンが変わり，環境によっては従来の ICM より早く学習することができる可能性がある．

色のネガポジ反転の変換処理では，エージェントが選択した行動は変換前と変わらないため，常に $\alpha(a_t) = a_t$ であるが，左右反転の変換処理では左右方向に関係する行動のみ入れ替わることになる．

5. 実装

本研究の実験は，機械学習用ソフトウェアライブラリである Tensorflow[6] を使用して行った．また，Doom の環境は研究用プラットフォームである Viz-Doom[7] を使用した．提案手法および実験のプログラムは，Pathak らが公開している ICM の Python プログラムコード [8] に変更を加える形で作成した．

6. 実験

実験環境は Viz-Doom の DoomMyWayHome-v0 を使用する．これは，強化学習プラットフォームの OpenAI Gym[9] の一種として利用することができる．この問題設定では，エージェントはマップ上の一箇所に配置されたベストへ最短経路を通って到達する行動の方策を獲得することが目標となる．エージェントは環境から図 2 のようなプレイヤーの主観視点 RGB 画像を受け取る．Mnih らの研究 [3][10] のように，受け取った画像は図 3 のようにグレースケールになり，42 × 42 ピクセルにリサイズされる．更に，時間的依存をモデル化するため，直前 3 フレームまでの合計 4 フレームをセットとして状態 s_t とされる．エージェントは環境から状態 s_t を受け取った際，4 つの行動 (何もしない，前に進む，右に進む，左に進む) のうちの 1 つを選択し，行動 a_t として環境に返す．訓練の高速化のため，エージェントは 4 ステップごとに行動を選択し，4 ステップの間は同じ行動を繰り返す．目標となるベストにエージェントが接触した場合，環境から報酬 1 が与えられ，それ以外の場合は負の報酬 -0.0001 が与えられる．エージェントが行動を開始してから，ベストを発見するか 2100 ステップが経過するまでを 1 エピソードとし，エージェントはエピソードの開始時に初期地点に再配置される．

行動の方策の学習手法には A3C を用いる．本実験では 20 のワーカーが非同期的にエージェントを訓練し，ADAM オプティマイザーを使用した．状態 s_t はそれぞ

図 2. ゲーム画面

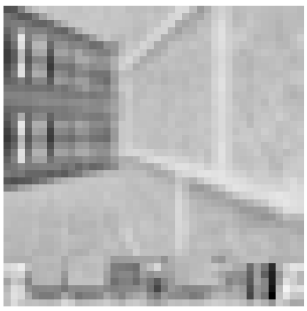

図 3. 前処理された状態

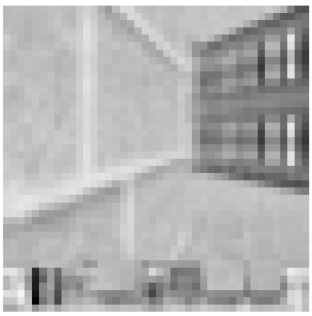

図 4. 左右反転した状態

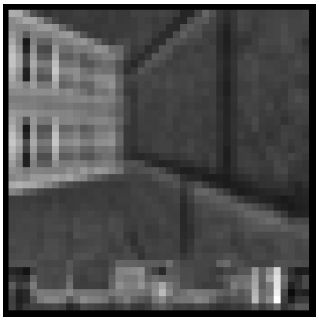

図 5. ネガポジ反転した状態

れ 32 のフィルター，3 × 3 のカーネル，2 のストライド，1 のパディングを持つ畳み込み層を 4 つ通過する．ELU[11] はそれぞれの畳み込み層の後に使用される．最後の畳み込み層の出力は LSTM に 256 ユニットで与えられる．

ICM の逆モデルも方策を学習するネットワークと同じく，それぞれ 32 のフィルター，3 × 3 のカーネル，2 のストライド，1 のパディングを持つ畳み込み層 4 つを通過し，入力状態 s_t を状態の特徴ベクトル $\phi(s_t)$ に写像する．ELU はそれぞれの畳み込み層の後に使用される．最後の畳み込み層の出力は 288 ユニットである．$\phi(s_t)$ と $\phi(s_{t+1})$ は 1 つの特徴ベクトルに連結し，入力として 256 ユニットの全結合層を通り，4 種類の行動のうちの 1 つを予測する 4 ユニットの全結合層から出力される．

順モデルは状態の特徴量 $\phi(s_t)$ とエージェントの取った行動 a_t を連結し，それぞれ 255 と 288 ユニットの 2 つの全結合層を通過させる．パラメータは，$\beta = 0.2$, $\lambda = 0.1$ である．また，学習率は 10^{-3} である．左右反転の変換処理を用いた ICM では，エージェントが選択した行動を変換する際に「右に進む」と「左に進む」が入れ替わる．左右反転処理，ネガポジ反転処理を行った状態を図 4，図 5 に示す．

6.1. 学習速度の比較

図 6 のように，9 の部屋と廊下で構成されたマップで従来の ICM+A3C，左右反転を用いた ICM+A3C，ネガポジ反転を用いた ICM+A3C の 3 つの手法でエージェントをそれぞれ訓練する．それぞれの手法は 20 のワーカーの探索ステップ数の合計が 10,000,000 ステップになるまで訓練を続けた．このマップは，Pathak らの実験で使用されたものと同じものである．マップの壁は部屋ごとに異なるテクスチャが貼られている．エピソードの開始時にエージェントが配置される初期地点はマップ全体に分布している 17 の地点から毎回ランダムで選ばれる．

3 つの手法が訓練の進行とともに各ステップで獲得した外部報酬の推移は図 7 のようになる．この環境の訓練において，エージェントが行動を選択した際，目標状態に到達していなければ環境から -0.0001 の報酬を受け取り，到達していれば 1 の報酬を受け取る．そのため，1 エピソードの中で獲得できる報酬の合計は，エピソードの終了条件である 2100 ステップの経過までに目標状態に到達できない訓練の初期段階では 0 を下回り，目標状態に到達できるようになってからはより短い経路で到達するほど 1 に近い値となる．図 7 に示されるとおり，全ての手法で報酬の値が収束しており，訓練の終了までに目標地点へ到達するための最適な方策を学習していることが分かる．従来の ICM+A3C に対して，ネガポジ反転を用いた ICM+A3C は訓練のより早い段階で報酬の値が収束している．一方，左右反転を用いた ICM+A3C は従来の ICM+A3C よりも報酬の値の収束が遅くなってしまっている．

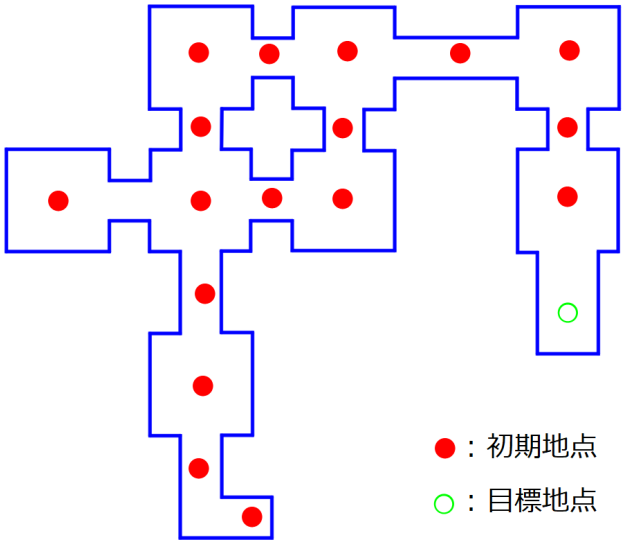

図 6. Doom のマップ

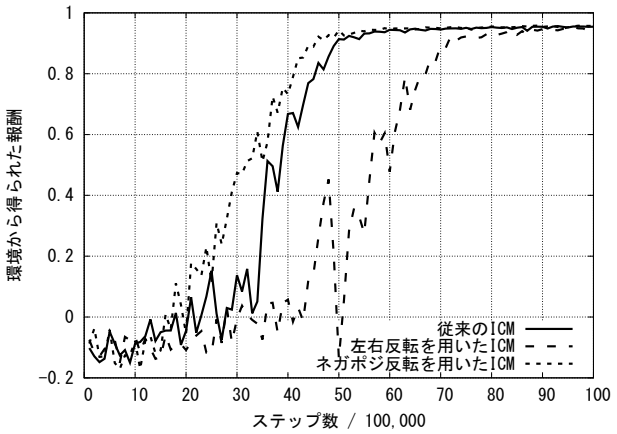

図 7. 獲得した外部報酬の推移

6.2. 汎化性能の評価

6.1 節の実験と同じマップで 3 つの手法を訓練し，学習内容を保持したままエージェントに新しいマップ (図

8) を訓練させる．これにより，各手法でエージェントが獲得した方策が，別の環境に移った際にどれだけ役立つのかを評価する．3 つの手法を図 6 のマップで事前に 10,000,000 グローバルステップ訓練する．事前訓練をどの手法で行っていても，新しいマップの訓練は従来の ICM+A3C で行う．また，事前訓練を行なった手法と比較するため，従来の ICM+A3C を用いて事前に訓練することなく新しいマップを一から訓練させる手法を実験に追加する．新しいマップは構造と壁のテクスチャが異なり，図 6 のマップと同じく 17 の初期地点が全体に分布している．

4 つの手法が訓練によって環境から獲得した外部報酬の推移は図 9 のようになる．エージェントを事前訓練させた 3 つの手法全てが事前訓練をしない従来の ICM+A3C よりも早く報酬の値が収束した．このため，事前学習でエージェントが獲得した方策は新しいマップでも役立っていることがわかる．事前学習をさせた 3 つの手法の中では，従来の ICM+A3C が最も早く報酬の値が収束した．

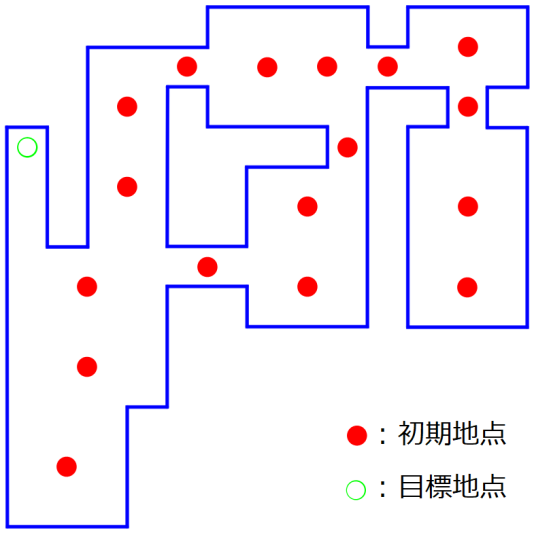

図 8. 汎化性能評価用のマップ

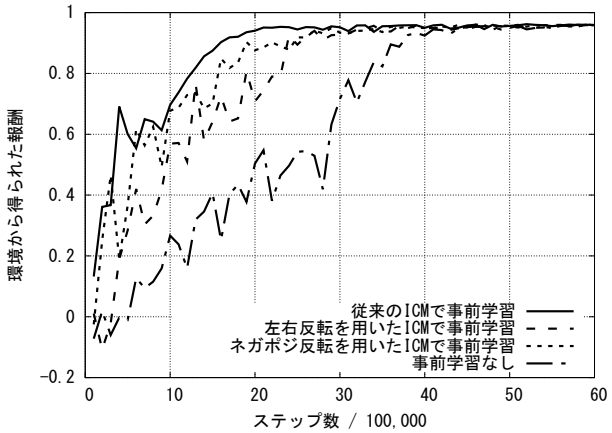

図 9. 事前学習をした ICM の比較

6.3. ランダムなテクスチャのマップの訓練

図 6 のマップは壁のテクスチャが部屋ごとに異なっており，同じ部屋の中ではテクスチャが統一されている．また，部屋と部屋をつなぐ廊下のテクスチャはマップ全体で共通である．

この環境のテクスチャの規則性が，ICM を組み合わせた強化学習の各手法に与える影響を調べるために，マップのテクスチャをランダムに配置した環境で訓練を行う．実験に使用するマップの構造は図 6 と同様で，元々各部屋と廊下に使用されていた 9 種類の壁のテクスチャを部屋か廊下かに関係なく壁面ごとにランダムに 1 つ選択して配置した．マップのテクスチャ以外の条件は 6.1 節の実験から変更しない．

各手法の訓練中に環境から与えられた報酬の値の推移は図 10 のようになる．この実験では，6.1 節の実験と異なり，従来の ICM+A3C よりも変換画像を用いた 2 つの手法の方が環境から得られた報酬の値が早く収束した．

それぞれの手法について，図 6 のマップで訓練を行なった結果とランダムなテクスチャのマップで訓練を行った結果を比較すると，従来の ICM+A3C は部屋ごとにテクスチャが決まっているマップの訓練に対してランダムなテクスチャのマップの訓練の方が報酬の値が収束するのが遅かった．左右反転を用いた ICM+A3C は従来の ICM+A3C とは逆に，部屋ごとにテクスチャが決まっているマップの訓練に対してランダムなテクスチャのマップの訓練の方が報酬の値が収束するのが早くなった．

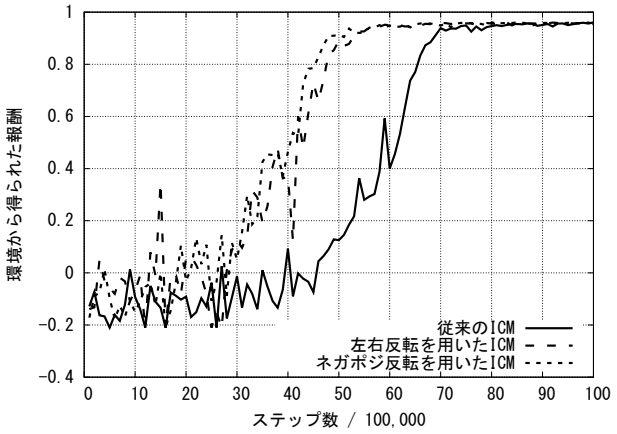

図 10. ランダムなテクスチャのマップでの訓練

7. 議論

6.1 節の実験では，状態のネガポジ反転を用いた ICM で訓練を行ったエージェントは，従来の ICM で訓練を行ったエージェントよりも環境から与えられる報酬の値が早い段階で最適化された一方で，左右反転を用いた ICM で訓練を行ったエージェントは他 2 つの手法に比べて大幅に報酬の値の収束が遅くなってしまった．そこで，この 2 つの手法の違いを考える．

ネガポジ反転を用いた ICM では，1 ステップの訓練で逆モデルは明度の異なる 2 通りの状態から同じ行動 a_t を予測できるように訓練されることになる．そのため，ネガポジ反転を用いた ICM の逆モデルから抽出される状態の特徴量 $\phi(s_t)$，$\phi(s_{t+1})$ は他の手法のものに比べて状態の明度に依らない特徴を早期に獲得することで，訓練の早期から壁の明度を学習から排除できるようになり，他の手法に対して優位となった可能性がある．

一方，左右反転を用いた ICM では，マップ上の部屋の特定の 1 つの中でエージェントが状態 s_t を受け取っ

たと仮定した場合，その状態を左右反転させた $\sigma(s_t)$ に近似できる環境中の状態は，壁のテクスチャが同じである，エージェントのいる部屋の中の状態ということになる．そのため，変換された状態を逆モデルの訓練に使用してもエージェントが現在いる部屋の特徴しか抽出できないので他の手法よりも最適な方策へ収束するのが遅くなってしまったのではないだろうか．

6.2 節の実験では，事前学習をしていた手法のエージェントのうち，最も報酬の値の収束が早かったのが従来の ICM+A3C であった．事前学習に使ったマップは，同じ大きさの部屋がつながった構造であり，テクスチャは部屋ごとに定まっていた．そのため，環境中の状態に変換処理を加えても，ほとんどは環境中の別の状態に近似できた．しかし，新しいマップは構造が複雑かつ対称性がなく，更にテクスチャも部屋で統一されている部分もあればそうでない部分もあるため，変換された状態が環境中に存在せず，変換処理を用いる ICM の手法は効率的な学習ができなかったのではないかと考えられる．

6.3 節の実験では，6.1 節の実験で使用した図 6 のマップを改変し，訓練を行うマップの部屋ごとに固有の壁のテクスチャを持つという規則性を取り除いた．その結果，左右反転を用いた ICM+A3C は，ランダムなテクスチャのマップでは従来の ICM+A3C よりも収束が早かった．

左右反転を用いた ICM+A3C でランダムなテクスチャのマップを探索した場合，ICM の逆モデルは状態 s_t，s_{t+1} を学習に使用すると同時に左右反転された状態 $\sigma(s_t)$，$\sigma(s_{t+1})$ を学習に使用するが，これらの変換された状態は画像中に複数の壁が写っている場合，壁面ごとに異なるテクスチャが使用されている．そのため，部屋ごとにテクスチャが決まっているマップと比較して変換された状態 $\sigma(s_t)$，$\sigma(s_{t+1})$ が状態 s_t，s_{t+1} とほとんど差異がないということが起きにくい．よって，逆モデルはより多様な状態をネットワークに学習させることができるようになり，学習に関係のある状態の特徴量 $\phi(s_t)$，$\phi(s_{t+1})$ を抽出するための学習をしやすいのではないかと考える．その結果，左右反転を用いた ICM+A3C でランダムなテクスチャのマップを訓練した結果は，部屋ごとにテクスチャが決まっているマップよりも報酬の値の収束が早くなった可能性がある．

一方，順モデルはランダムなテクスチャのマップでの訓練において，左右反転された状態の特徴量と行動 $\phi(\sigma(s_t))$，$\alpha(a_t)$ を入力として用いるが，エージェントが廊下などの通路を通った場合，ランダムなテクスチャのマップでは通路の左右の壁はそれぞれ異なるテクスチャが使用されており，その時の状態を左右反転した画像は，エージェントが現在通り抜けようとしている通路を反対方向から通り抜けようとしているときの状態に近似できる．順モデルは通路を通り抜けようとする行動に対する訓練を行い，同時に近似された通路を反対方向から通り抜けようとする行動に対する訓練も行うため，エージェントは通ってきた通路を逆走するような行動に対する内部報酬が小さくなり，そのような行動を取りにくくなると考えられる．そのこともランダムなテクスチャのマップでは左右反転を用いた ICM+A3C の学習の成績が良いことの助けになっていると考えられる．

部屋ごとにテクスチャが決まっているマップの探索では，エージェントがどの部屋にいるのかの情報を，状態の画像中に部屋の壁のいずれかが写っていることによって取得できたが，ランダムなテクスチャのマップでは，ある部屋で使用されていたテクスチャが別の部屋でも使用されている可能性があるため，部屋のどの壁面にどのテクスチャが使用されているのか，部屋につながる廊下のテクスチャはどれであるかなど現在エージェントがいる部屋を特定するために必要な情報が多くなっている．そのため従来の ICM+A3C は，環境から受け取った状態と，それを変換して近似した状態を同時に訓練に使用できる，変換された画像を用いた ICM+A3C に比べて，学習の進行が遅くなってしまったと考えられる．

8. おわりに

本研究では，ICM と組み合わせた強化学習のエージェントが環境から状態を受け取り，行動を選択して ICM のニューラルネットワークを訓練する際に，状態や行動をそのまま訓練させるだけでなく，変換処理を加えたものも訓練に使用することで学習速度を向上させる手法を提案した．実験を行なった結果，部屋ごとにテクスチャが決まっているマップにおいて，ネガポジ反転を用いる ICM で訓練したエージェントは従来の ICM で訓練したエージェントよりも早く最適な方策を獲得したが，左右反転を用いる ICM で訓練したエージェントは従来の ICM よりも最適な方策を獲得するのが遅かった．また，事前学習を行った上で別の環境で訓練する実験では，変換画像を用いる ICM で訓練したエージェントは，従来の ICM で訓練したエージェントに比べて汎化能力が低くなっていることが分かった．壁面ごとにランダムにテクスチャを使用したマップの訓練では，左右反転を用いた ICM で訓練したエージェントとネガポジ反転を用いた ICM で訓練したエージェントはどちらも従来の ICM で訓練したエージェントよりも早く最適な方策を獲得することができた．

今後の課題は，環境や変換処理の条件を変更すると ICM と組み合わせた強化学習の性能にどのような影響があるのかを調査することである．

参考文献

[1] R. S. Sutton, “強化学習,” 2000.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, “Playing atari with deep reinforcement learning,” *arXiv preprint arXiv:1312.5602*, 2013.

[3] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, “Asynchronous methods for deep reinforcement learning,” in *International conference on machine learning*, 2016, pp. 1928–1937.

[4] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell, “Curiosity-driven exploration by self-supervised prediction,” *arXiv preprint arXiv:1705.05363*, 2017.

[5] 浅川伸一, “深層学習をめぐる最近の熱狂,” 基礎心理学研究, vol. 35, no. 2, pp. 149–162, 2017.

[6] “Tensorflow,” https://www.tensorflow.org/.

[7] M. Kempka, M. Wydmuch, G. Runc, J. Toczek, and W. Jaśkowski, “Vizdoom: A doom-based ai research platform for visual reinforcement learning,” in *Computational Intelligence and Games (CIG), 2016 IEEE Conference on*. IEEE, 2016, pp. 1–8.

[8] “Github - pathak22/noreward-rl: [icml 2017] tensorflow code for curiosity-driven exploration for deep reinforcement learning,” https://github.com/pathak22/noreward-rl.

[9] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, “Openai gym,” *arXiv preprint arXiv:1606.01540*, 2016.

[10] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski *et al.*, “Human-level control through deep reinforcement learning,” *Nature*, vol. 518, no. 7540, p. 529, 2015.

[11] D.-A. Clevert, T. Unterthiner, and S. Hochreiter, “Fast and accurate deep network learning by exponential linear units (elus),” *arXiv preprint arXiv:1511.07289*, 2015.